PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2004-030093
(43)Date of publication of application : 29.01.2004

(51)Int.Cl. G06F 17/30
C12N 15/00
// G01N 33/53
G01N 37/00

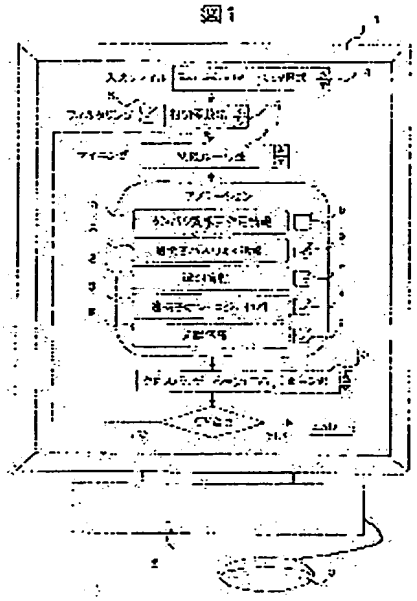
(21)Application number : 2002-183810
(22)Date of filing : 25.06.2002

(71)Applicant : HITACHI LTD
(72)Inventor : TOMITA HIROYUKI
MAKI HIDEYUKI
MORITA TOYOHISA
SORIN SABAU
TANIGAWA KOJI

(54) METHOD FOR ANALYZING GENE EXPRESSION DATA

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a method and equipment for knowledge search based on gene expression data (also called a gene expression profile) using a DNA micro-array or the like.
SOLUTION: The knowledge search is done through: a process receiving the gene expression data; a process receiving class information; a process extracting a genetic group related to class classification by using a data mining technique; a process executing annotation with respect to the genetic group; a process extracting the common rule of the genetic group related to the class classification based on the genetic annotation; and a process executing the data mining using constraint conditions based on the common rule.

図1



LEGAL STATUS
[Date of request for examination]
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

Japanese Laid-Open Patent Publication
No. 30093/2004 (*Tokukai* 2004-30093)

A. Relevance of the Above-identified Document

The following is a partial English translation of exemplary portions of non-English language information that may be relevant to the issue of patentability of the claims of the present application.

B. Translation of the Relevant Passages of the Document

See also the attached English Abstract.

[INDUSTRIAL FIELD OF THE INVENTION]

The present invention relates to a method and an apparatus for finding information based on gene expression data (also referred to as "gene expression profile") using a DNA microarray or the like.

[0002]

[PRIOR ART]

Research is underway to find gene functions through gene expression profile analysis and obtain information useful for drug development, pharmacology, toxicology, and diagnosis. For the analysis of DNA chip data, the following techniques have been used, for example: statistical analyses such as correlation analysis, principal component analysis, and analysis of variance; clustering such as k-mean clustering, hierarchical clustering, and self-organizing map; and classification algorithm such as nearest neighbor, discrimination analysis, support vector machine, neural network, and genetic algorithm (see Laura J. van't Veer et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415, pp. 530-536 (2002); Scott L. Pomeroy et al., Prediction of central nervous

system embryonal tumor outcome based on gene expression, Nature 415, pp. 436-442 (2002)).

[0003]

[PROBLEMS TO BE SOLVED BY THE INVENTION]

However, there is no established means that can be used to simultaneously analyze large numbers of genes and obtain information. As to the information obtained by the analysis, it greatly depends on the knowledge of the analyzer. As such, the same information cannot always be obtained between different analyzers. Further, the amount of data obtained from the DNA chip is so large that it exceeds the analyzing capability of humans.

[0004]

An object of the present invention is to provide a method and an apparatus for analyzing a gene expression profile set. More specifically, the invention provides a technique for extracting, from a gene expression profile set, a group of genes that are useful for drug development, pharmacology, toxicology, and diagnosis, and a technique for finding laws common to such groups of useful genes.

[0005]

[MEANS TO SOLVE THE PROBLEMS]

In order to extract a group of useful genes from a gene expression profile set, and find laws common to such groups of useful genes, the invention performs the steps of (1) receiving gene expression data, (2) receiving class information, (3) extracting groups of genes associated with class classification, using a data mining method, (4) performing annotation on the groups of genes, (5) finding laws common to the groups of genes associated with class classification, based on gene annotation, and (6) performing data mining using restricting conditions that are based on the common laws. It is important that the

procedure from step (3) to step (6) be repeated. Conventional information systems employ steps (1) through (4). Conventionally, the groups of genes associated with class classification, and their annotations are presented to an analyzer in the form of, for example, a list or a graphical interface. The subsequent step of obtaining information relied on knowledge and intuition of the analyzer. A problem of this approach, however, is that analyzers have different knowledge levels. Another problem is that the amount of information obtained from the DNA chip gene expression profile is so large that the analysis takes a long time. Accordingly, there is a need for an information system for assisting an analyzer in obtaining information.

(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2004-30093
(P2004-30093A)
(43)公開日　平成16年1月29日(2004.1.29)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	170F	5B075
C12N 15/00	G06F 17/30	220Z	
// G01N 33/53	C12N 15/00	Z	
G01N 37/00	G01N 33/53	M	
	G01N 37/00	102	

審査請求 未請求　請求項の数 17　OL　（全 29 頁）

(21)出願番号　特願2002-183810(P2002-183810)
(22)出願日　平成14年6月25日(2002.6.25)

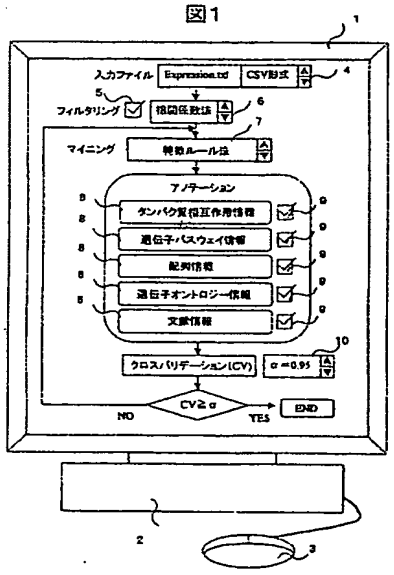
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(74)代理人　100075096
弁理士　作田　康夫
(72)発明者　富田　裕之
東京都千代田区神田駿河台四丁目６番地
株式会社日立製作所ライフサイエンス推進事業部内
(72)発明者　牧　秀行
神奈川県川崎市麻生区王禅寺１０９９番地
株式会社日立製作所システム開発研究所内

最終頁に続く

(54)【発明の名称】遺伝子発現データ解析方法

(57)【要約】
【課題】本発明の目的は、ＤＮＡマイクロアレイ等を用いた遺伝子発現データ（遺伝子発現プロファイルとも言う）に基づく、知識探索の方法および装置を提供することにある。
【解決手段】遺伝子発現データを受け取る工程、クラス情報を受け取る工程、データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程、前記遺伝子群にアノテーションを行う工程、遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程、前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程を行うことで知識探索を行う。
【効果】遺伝子発現データからメタ知識を得ることができる。
【選択図】　図１

図1



【特許請求の範囲】
【請求項１】
遺伝子発現データを受け取る工程と、クラス情報を受け取る工程と、データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程と、前記遺伝子群に注釈付け（アノテーション）を行う工程と、前記遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程と、前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程を有することを特徴とする遺伝子発現データ解析方法。
【請求項２】
前記遺伝子発現データを受け取る工程の後、前記データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程の前に、遺伝子発現データを正規化する工程と、実験ばらつきを考慮して遺伝子発現データをフィルタリングする工程とを有することを特徴とする請求項１記載の遺伝子発現データ解析方法
【請求項３】
前記クラス分類に関連する遺伝子群を抽出する工程において抽出された遺伝子群を用いてクロスバリデーションを行うことにより、該データマイニング手法を用いたクラス分類の正解率もしくはエラー率を比較する工程を、更に有することを特徴とする請求項１または２記載の遺伝子発現データ解析方法
【請求項４】
前記データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程と、前記遺伝子群にアノテーションを行う工程と、前記遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程、前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程を、複数回繰り返すことを特徴とする請求項１乃至３何れかに記載の遺伝子発現データ解析方法。
【請求項５】
前記クロスバリデーションにより得られる、該データマイニング手法を用いたクラス分類の正解率が、既定のしきい値を超えるまで、データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程と、前記遺伝子群にアノテーションを行う工程と、遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程と、前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程とを繰り返すことを特徴とする請求項３記載の遺伝子発現データ解析方法。
【請求項６】
前記データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程とは、遺伝子データから、グループ間二乗和とグループ内二乗和の比率（ＢＳＳ／ＷＳＳ比）もしくは特徴ルール法での正解率が、上位１０個から２００個の遺伝子を抽出する工程であることを特徴とする請求項１乃至５何れかに記載の遺伝子発現データ解析方法。
【請求項７】
前記遺伝子アノテーションに基づきクラス分類に関連する遺伝子群の共通規則を抽出する工程は、前記クラス分類に関連する遺伝子群を抽出する工程により抽出された個々の遺伝子について、遺伝子間の相互関係である二項関係、パスウェイ、ゲノム、階層構造、ネットワーク関係のいずれか１つもしくは複数に属する共通ルールを検索する工程であることを特徴とする請求項１乃至５何れかに記載の遺伝子発現データ解析方法。
【請求項８】
前記相互関係が二項関係のリガンドーレセプターの関係であり、前記抽出された遺伝子がリガンド遺伝子であれば、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件はレセプター遺伝子もしくはリガンド遺伝子とレセプター遺伝子の組を含む条件とし、前記抽出遺伝子がレセプター遺伝子であれば、前記拘束条件をリガンド遺伝子もしくはリガンド遺伝子とレセプター遺伝子の組を含む条件とすることを特徴とする請求項７記載の遺伝子発現データ解析方法。
【請求項９】
前記相互関係がパスウェイであり、前記抽出遺伝子がパスウェイＰＡ上にあり、クラス分

類に関連する遺伝子群の共通規則をパスウェイＰＡ上の上流遺伝子、下流遺伝子、パスウェイＰＡと相関するパスウェイＰＢ上の遺伝子、もしくは前記いずれかの遺伝子の組を含む条件とすることを特徴とする請求項７記載の遺伝子発現データ解析方法。

【請求項１０】

前記相互関係がゲノムであり、前記抽出遺伝子が染色体ＣＡ上にあり、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を染色体ＣＡ上の隣接遺伝子、もしくは前記抽出遺伝子と前記の隣接遺伝子との組を含む条件とすることを特徴とする請求項７記載の遺伝子発現データ解析方法。

【請求項１１】

前記相互関係が階層構造でかつオントロジーであり、前記抽出遺伝子のオントロジーがＯＡであり、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件をオントロジーＯＡの上階層のオントロジーを有する遺伝子、もしくは前記抽出遺伝子と前記の上階層のオントロジーを有する遺伝子との組を含む条件とすることを特徴とする請求項７記載の遺伝子発現データ解析方法。

【請求項１２】

前記相互関係が階層構造でかつ酵素（ＥＣ）であり、前記抽出遺伝子のＥＣ番号（Ｅｎｚｙｍｅ　Ｃｏｍｍｉｓｓｉｏｎ）がＥＣＡであれば、請求項１記載の、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を酵素ＥＣＡの上階層に属する遺伝子、酵素ＥＣＡと同一グループに属する遺伝子もしくは前記いずれかの遺伝子の組を含む条件とすることを特徴とする請求項７記載の遺伝子発現データ解析方法。

【請求項１３】

前記相互関係が階層構造でかつスーパーファミリーであり、前記抽出遺伝子のスーパーファミリーがＳＦＡであり、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を同一スーパーファミリーＳＦＡに属する遺伝子、もしくは前記抽出遺伝子と前記の同一スーパーファミリーの属する遺伝子との組を含む条件とすることを特徴とする請求項７記載の遺伝子発現データ解析方法。

【請求項１４】

前記相互関係がネットワークでかつ文献情報であり、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を、文献情報により、前記抽出遺伝子との関連が予想される遺伝子、もしくは前記抽出遺伝子と前記の文献情報から関連が予想される遺伝子との組を含む条件とすることを特徴とする請求項７記載の遺伝子発現データ解析方法。

【請求項１５】

前記相互関係がネットワークでかつ蛋白質相互作用である、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を、蛋白質相互作用により、前記抽出遺伝子との関連が予想される遺伝子、もしくは前記抽出遺伝子と前記の蛋白質相互作用から関連が予想される遺伝子との組を含む条件とすることを特徴とする請求項７記載の遺伝子発現データ解析方法

【請求項１６】

前記相互関係がネットワークでかつ代謝経路情報であり、クラス分類に関連する遺伝子群の共通規則に基づく拘束条件を、代謝経路情報により、前記抽出遺伝子との関連が予想される遺伝子、もしくは前記抽出遺伝子と前記の代謝経路情報から関連が予想される遺伝子との組を含む条件とすることを特徴とする請求項７記載の遺伝子発現データ解析方法

【請求項１７】

前記遺伝子アノテーションに基づきクラス分類に関連する遺伝子群の共通規則を抽出する工程は、クラス分類に関連する遺伝子群を抽出する工程により抽出された個々の遺伝子について、遺伝子全体像（ゲノム）、転写産物全体像（トランスクリプトーム）、タンパク質全体像（プロテオーム）、酵素全体像（エンザイモーム）、代謝全体像（メタボローム）、相互作用全体像（インタラクトーム）、時間的空間的局在全体像（ローカリゾーム）、表現型全体像（フェノーム）いずれか１つ以上の全体像内における相互作用内の共通ルールを検索する工程であることを特徴とする請求項１記載の遺伝子発現データ解析方法。

【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、ＤＮＡマイクロアレイ等を用いた遺伝子発現データ（遺伝子発現プロファイルとも言う）に基づく、知識探索の方法および装置に関する。
【０００２】
【従来の技術】
遺伝子発現プロファイルを解析することで遺伝子機能を明らかにし、創薬、薬理学、毒性学、診断に供する知見を得るための研究がなされている。例えば、相関分析（Ｃｏｒｒｅｌａｔｉｏｎ Ａｎａｌｙｓｉｓ）、主因子分析（Ｐｒｉｎｃｉｐａｌ Ｃｏｍｐｏｎｅｎｔ Ａｎａｌｙｓｉｓ）、分散分析（Ａｎａｌｙｓｉｓ ｏｆ Ｖａｒｉａｎｃｅ）などの統計解析、ｋ平均クラスタリング（ｋ－ｍｅａｎ Ｃｌｕｓｔｅｒｉｎｇ）、階層クラスタリング（Ｈｉｅｒａｒｃｈｉｃａｌ Ｃｌｕｓｔｅｒｉｎｇ）、自己組織化マップ（Ｓｅｌｆ－ｏｒｇａｎｉｚｉｎｇ Ｍａｐ）などのクラスタリング、最短近傍法（Ｎｅａｒｅｓｔ Ｎｅｉｇｈｂｏｒ）、判別分析（Ｄｉｓｃｒｉｍｉｎａｎｔ Ａｎａｌｙｓｉｓ）、サポートベクターマシン（Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅ）、ニューラルネットワーク（Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋ）、遺伝的アルゴリズム（Ｇｅｎｅｔｉｃ Ａｌｇｏｒｉｓｍ）などの分類アルゴリズムをＤＮＡチップデータの解析に適用した例がある。Ｌａｕｒａ Ｊ. ｖａｎ’ｔ Ｖｅｅｒら、Ｇｅｎｅ ｅｘｐｒｅｓｓｉｏｎ ｐｒｏｆｉｌｉｎｇ ｐｒｅｄｉｃｔｓ ｃｌｉｎｉｃａｌ ｏｕｔｃｏｍｅ ｏｆ ｂｒｅａｓｔ ｃａｎｃｅｒ. Ｎａｔｕｒｅ ４１５, ｐｐ. ５３０－５３６ （２００２）や、Ｓｃｏｔｔ Ｌ. Ｐｏｍｅｒｏｙら、Ｐｒｅｄｉｃｔｉｏｎ ｏｆ ｃｅｎｔｒａｌ ｎｅｒｖｏｕｓ ｓｙｓｔｅｍｅｍｂｒｙｏｎａｌ ｔｕｍｏｒ ｏｕｔｃｏｍｅ ｂａｓｅｄ ｏｎ ｇｅｎｅ ｅｘｐｒｅｓｓｉｏｎ. Ｎａｔｕｒｅ ４１５, ｐｐ. ４３６－４４２ （２００２）を参照。
【０００３】
【発明が解決しようとする課題】
しかしながら、多数の遺伝子の発現を同時に解析して知識を獲得する手段については、いまだに確立されたとはいえない状況である。特に知識獲得については、解析者の知識に依存することが大きいので、解析者によって得られる知識に差があること、またＤＮＡチップから得られるデータ量が膨大であり、そもそも人間の解析能力を超えているという問題があった。
【０００４】
本発明の目的は、遺伝子発現プロファイルのセットを解析するための方法および装置を提供することである。更に詳しくは、遺伝子発現プロファイルのセットから、創薬、薬理学、毒性学、診断に有用な遺伝子群の抽出およびそれら有用遺伝子群に共通する規則を抽出する技術を提供することを目的とする。
【０００５】
【課題を解決するための手段】
遺伝子発現プロファイルのセットから有用な遺伝子群の抽出およびそれら有用遺伝子群に共通する規則を抽出するために、（１）遺伝子発現データを受け取る工程、（２）クラス情報を受け取る工程、（３）データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程、（４）前記遺伝子群にアノテーションを行う工程、（５）遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程、（６）前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程を行う。また前記（３）から（６）の工程を反復することが重要である。従来の情報システムは工程１から４までを行っていた。クラス分類に関連する遺伝子群とそのアノテーションについて、例えば一覧表、あるいはグラフィカルユーザーインターフェースを用いて、解析者に提示するに留めていた。その後の知識獲得の工程は、解析者の知識と直感に頼っていた。しかし、解析者ごとの知識レベルが異なること、そもそもＤＮＡチップ遺伝子発現プロファイ

ルデータから得られる情報が膨大であるので、解析に多大の時間がかかることが問題となっている。そのため知識獲得においても、解析者を支援する情報システムが必要とされている。

【０００６】

遺伝子発現解析を除く分野でのデータマイニングは、一般にインスタンス数＞＞アトリビュート数の関係にある。例えばデパートの顧客データを例にすると、インスタンス数（顧客数）イコール数千から数万、アトリビュート数（性別、年齢、年収など）イコールたかだか百個であり、上記の関係が成り立つ。しかしＤＮＡチップでは反対に、インスタンス数（サンプル数：数個から数十）に対し、アトリビュート数（遺伝子数：数千から数万）が桁違いに大きいことが特徴である。この種の問題では、「高度なマイニング手法を用いることなく、既存のサンプルに対してある程度正確な予測ができる」一方で「新規なサンプルに対しては予測に失敗することが多い」ことが知られている。この理由は、例えば１０個の遺伝子で説明できなければ、１００個、２００個と遺伝子を増やすことでいつかは説明ができるためである。また実験誤差をあらかじめ除くことが難しいのでアトリビュート値の信頼性が低いこと、そして誤り値（Ｉｎａｃｃｕｒａｔｅ ｖａｌｕｅｓ）が少しでも含まれているとマイニング結果が大きく変わってしまう（ロバストでない）ことも知られている。このため、適切なデータを選択すること（Ｄａｔａ Ｓｅｌｅｃｔｉｏｎ）、および解析に悪影響を与える誤り値を除くこと（Ｄａｔａ Ｃｌｅａｎｓｉｎｇ）などの前処理が重要となる。現在、Ｄａｔａ Ｓｅｌｅｃｔｉｏｎ、Ｄａｔａ Ｃｌｅａｎｓｉｎｇの確立した方法はないが、後のＫｎｏｗｌｅｄｇｅ Ｄｉｓｃｏｖｅｒｙの成否を決定的に左右すると考えられる。すなわちＤＮＡチップデータ等の遺伝子発現データの性質は、（１）アトリビュート数＞＞インスタンス数、（２）アトリビュート値の信頼性が低い、（３）誤り値に対しロバストでないので、Ｄａｔａ Ｓｅｌｅｃｔｉｏｎ、Ｄａｔａ Ｃｌｅａｎｓｉｎｇは極めて重要であるという３つの性質である。以後、本願明細書の記載では、Ｄａｔａ Ｓｅｌｅｃｔｉｏｎ、Ｄａｔａ Ｃｌｅａｎｓｉｎｇをまとめてフィルタリングと呼ぶ。

【０００７】

遺伝子発現データとは、例えばＤＮＡチップ法（ＤＮＡ Ｃｈｉｐ）、ディファレンシャルディスプレイ法（Ｄｉｆｆｅｒｅｎｔｉａｌ Ｄｉｓｐｌａｙ）、定量的ＰＣＲ法（Ｑｕａｎｔｉｔａｔｉｖｅ ＰＣＲ）、ＳＡＧＥ（Ｓｅｒｉａｌ Ａｎａｌｙｓｉｓ ｏｆ Ｇｅｎｅ Ｅｘｐｒｅｓｓｉｏｎ）法、プロテインチップ法（Ｐｒｏｔｅｉｎ Ｃｈｉｐ）などの複数遺伝子もしくは蛋白質の発現変化を測定する方法により得られた複数遺伝子（あるいは蛋白質）に関する発現量、もしくは発現量同士の比率のことである。

【０００８】

クラス情報とは、例えばＤＮＡチップ法等で測定された対象を分類するための情報である。例えば測定対象サンプルが被検査者の血液である場合、その血液が患者由来もしくは健常人由来のいずれであるかを、患者由来であれば１、健常人由来であれば０と定義する。また病理知見等に基づき、がんなどの疾患の悪性度を０，１，２，３などと定義できる。培養細胞であれば、薬物投与前の細胞由来サンプルを０、薬物投与６時間後のサンプルを１、薬物投与１２時間後のサンプルを２、薬物投与２４時間後のサンプルを３などと定義できる。クラスが同一であれば、測定対象となる個人、個体が異なっても、同一の実験条件や性質、表現型（フェノタイプ）を有するものとする。

【０００９】

データマイニングとは、データベースから、興味ある規則性や因果関係を計算機で自動的に抽出する技術のことである。例えば決定木（Ｄｅｃｉｓｉｏｎ Ｔｒｅｅ）、ナイブベイズ（Ｎａｉｖｅ Ｂａｙｅｓ）、フルベイジアン（Ｆｕｌｌｙ Ｂａｙｅｓｉａｎ）、相関ルール（Ａｓｓｏｃｉａｔｉｏｎ ｒｕｌｅ）、特徴ルール（Ｃｈａｒａｃｔｅｒｉｓｔｉｃ ｒｕｌｅ）、ＥＭクラスタリング（ＥＭ Ｃｌｕｓｔｅｒｉｎｇ）、最短近傍法（Ｎｅａｒｅｓｔ Ｎｅｉｇｈｂｏｒ）、判別分析（Ｄｉｓｃｒｉｍｉｎａｎｔ Ａｎａｌｙｓｉｓ），サポートベクターマシン（Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉ

ｎｅ）、遺伝的アルゴリズム（Ｇｅｎｅｔｉｃ　Ａｌｇｏｒｉｓｍ）、線形回帰（ＬｉｎｅａｒＲｅｇｒｅｓｓｉｏｎ）のことである。もしくは上記方法を使用する上で、バギング（Ｂａｇｇｉｎｇ）、ブースティング（Ｂｏｏｓｔｉｎｇ）、スタッキング（Ｓｔａｃｋｉｎｇ）などの手法を併用することもある。

【００１０】

アノテーションとは注釈付けのことである。例えば塩基配列、遺伝子機能情報、疾患関連情報、公共データベースの該当ＩＤ、他種遺伝子間（例えばヒトとマウス）のホモログ情報、遺伝子ネットワーク情報、パスウェイ情報等である。

【００１１】

正規化（ノーマリゼーションとも言う）とは、実験を行った時間、場所、作業者が異なることから実験のバックグランドや、ノイズの程度が異なるので、その実験ごとのバックグランドやノイズの程度を揃える操作のことである。ＤＮＡチップであれば、チップごとに画像輝度（蛍光強度）が異なることがあるので、まずバックグランド輝度、例えば、本来プローブが存在しない部位の輝度を取得して、そのバックグランド輝度の平均値や中央値を、プローブ蛍光強度から差し引くことなどの操作のことである。

【００１２】

フィルタリングとは、前記のように適切なデータを選択し、解析に悪影響を与える誤り値を除く操作のことである。例えばしきい値を定める方法がある。具体的には、信号強度が小さい（例えば５００以下の）場合、しきい値イコール５００とし、５００以下の信号強度を５００にするかもしくはゼロにする操作のことである。

【００１３】

クロスバリデーション（Ｃｒｏｓｓ－Ｖａｌｉｄａｔｉｏｎ）とは、例えばテンフォールドクロスバリデーション（ｔｅｎｆｏｌｄ　ｃｒｏｓｓ－ｖａｌｉｄａｔｉｏｎ）、もしくはリーブワンアウトクロスバリデーション（Ｌｅａｖｅ－ｏｎｅ－ｏｕｔ　ｃｒｏｓｓ－ｖａｌｉｄａｔｉｏｎ）のことである。テンフォールドクロスバリデーションとは、データセットをランダムに１０個に等分割し、１０分の９個のデータでトレーニングし、残りの１０分の１個のデータでテストする合計１０回の試行の結果から、正解率（もしくはエラー率）を算出する方法である。リーブワンアウトクロスバリデーションとは、ｎ個のデータセットのうち、ｎ－１個のデータでトレーニングし、残りの１個のデータでテストする合計ｎ回の試行の結果から、正解率（もしくはエラー率）を算出する方法である。限られた数のデータセットから、各マイニング方法の正解率（もしくはエラー率）を比較し、どのマイニング方法が優れているのかを定量的に評価する方法である。

【００１４】

図１は、本発明のソフトウェアを実行するために利用され得るコンピュータシステムと全体の工程を表したインターフェースの一例を示す。まず遺伝子発現データ、クラス情報が格納されているファイル入力のためにファイル形式やファイル名を入力する。続いて、正規化やフィルタリング方法の選択をし、データマイニングを行う。データマイニングにより抽出された遺伝子群にアノテーションをほどこす際に、どのアノテーションを行うかを選択する。またデータマイニングにより抽出された遺伝子群に基づいてクロスバリデーションを計算し、正解率がしきい値（α；図１では０．９５）以上ならば、そのまま終了し、しきい値以下であれば、アノテーション結果を反映させた拘束条件のもとで次回のマイニングを行う。このマイニングは、正解率がしきい値を超えるまで自動的に反復して行える。

【００１５】

図２は、ＤＮＡチップの一般的な構造を示した図である。図１６にＤＮＡチップを用いた測定法のフローチャートを示す。まず支持体２４にＤＮＡプローブ２２を固定化する。続いて、測定対象サンプルから抽出した遺伝子断片を蛍光標識などで標識する。この蛍光標識された遺伝子２３を、ＤＮＡプローブ２２とハイブリダイズさせる。その後、蛍光標識由来の蛍光を検出器２１で検出する。この検出の結果、各ＤＮＡプローブ２２にハイブリダイズした蛍光標識された遺伝子２３の量が得られる。これを発現分布という。

【００１６】
図３は、遺伝子アノテーションを行うに際して考慮すべき、遺伝子全体像（ゲノム：Ｇｅｎｏｍｅ）、転写産物全体像（トランスクリプトーム：Ｔｒａｎｓｃｒｉｐｔｏｍｅ）、蛋白質全体像（プロテオーム：Ｐｒｏｔｅｏｍｅ）における遺伝子同士、転写産物同士、蛋白質同士、あるいは遺伝子と転写産物、転写産物と蛋白質、遺伝子と蛋白質の相互関係の例である。なお、オーム（ｏｍｅ）は全体あるいは全体像を意味する接尾語であり、遺伝子（Ｇｅｎｅ）の全体をゲノム（Ｇｅｎｏｍｅ）、転写産物（Ｔｒａｎｓｃｒｉｐｔ）の全体をトランスクリプトーム（Ｔｒａｎｓｃｒｉｐｔｏｍｅ）、蛋白質（Ｐｒｏｔｅｉｎ）の全体をプロテオーム（Ｐｒｏｔｅｏｍｅ）と呼ぶ。本願明細書の以下の記述では、遺伝子、転写産物、蛋白質を遺伝子等と呼ぶ。図３における白丸が個々の遺伝子等を指し、白丸と白丸とをつないだ線は、実験等により既知となっている相互作用、因果関係である。図３（Ａ）は遺伝子等に相互作用がない状態であり、独立と呼ばれる。但し将来、実験が行われることによってなんらかの相互作用、因果関係が発見される可能性はある。図３（Ｂ）は遺伝子等が１対１で相互作用する関係であり、二項関係と呼ばれる。例えば図４に示すように細胞表面に存在する受容体（レセプター）とその受容体に結合する結合体（リガンド）の関係が二項関係の一例である。図４（Ａ）はインターロイキン２（ＩＬ２）蛋白質とインターロイキン２受容体アルファ（ＩＬ２ＲＡ）、インターロイキン２（ＩＬ２）蛋白質とインターロイキン２受容体ベータ（ＩＬ２ＲＢ）、インターロイキン２（ＩＬ２）蛋白質とインターロイキン２受容体ガンマ（ＩＬ２ＲＧ）、図４（Ｂ）は形質転換成長因子ベータ１（ＴＧＦＢ１）と形質転換成長因子ベータ受容体１（ＴＧＦＢＲ１）、形質転換成長因子ベータ１（ＴＧＦＢ１）と形質転換成長因子ベータ受容体２（ＴＧＦＢＲ２）、形質転換成長因子ベータ１（ＴＧＦＢ１）と形質転換成長因子ベータ受容体３（ＴＧＦＢＲ３）、図４（Ｃ）はエリスロポエチン（ＥＰＯ）蛋白質とエリスロポエチン受容体（ＥＰＯＲ）の間のリガンドーレセプター関係である。またＤＮＡとＤＮＡ結合蛋白質も二項関係の一例である。ＤＮＡ結合蛋白質の例として転写因子、修復遺伝子などがある。図３（Ｃ）は道筋にそって遺伝子等同士が相互作用する関係でありパスウェイと呼ばれる。パスウェイ中には分岐は存在するものの、上流から下流へと一方向に相互作用が行われることが特徴である。パスウェイ上流とパスウェイ下流の遺伝子等の間には、因果関係が存在するともいえる。パスウェイの例として、図５に示すＭＡＰキナーゼ（Ｍｉｔｏｇｅｎ Ａｃｔｉｖａｔｅｄ Ｐｒｏｔｅｉｎ Ｋｉｎａｓｅ）パスウェイのように、細胞表面の受容体とリガンドが結合したことを起点として、細胞表面から細胞内の細胞核にいたるまで情報を伝達するパスウェイが知られている。個々の丸印は遺伝子を、丸印と丸印をつなぐ矢印は、その矢印の方向に情報伝達が行われることを意味する。例えば図５のＭｏｓ遺伝子からＭＥＫ遺伝子に、ＭＥＫ遺伝子からＥＲＫ遺伝子に情報が伝達される。パスウェイ情報としては、例えばパスウェイデータベース　（ｈｔｔｐ：／／ｗｗｗ．ｂｉｏｃａｒｔａ．ｃｏｍ／）を参照。

【００１７】
図３（Ｄ）はＤＮＡ塩基配列上の遺伝子の相互配置関係であり、本願明細書ではゲノムと呼ぶ。ヒト、マウス、ラット等の高等動物では染色体上に、酵母や細菌などでは環状ＤＮＡ上に遺伝子が配置されている。図６にゲノムの例を示す。ヒト第１３番染色体上の１３ｑ１２から１３ｑ１３領域の一部では、ＬＯＣ２２２４２８遺伝子からＬＯＣ１６０９７９遺伝子までが、図６のような位置に順番に存在している。ある疾患は染色体の一部の領域が欠失したり、増幅したりする結果、近傍にある遺伝子群が同時に欠失もしくは増幅することが原因で発症する。そのよう遺伝子増幅、遺伝子欠失が原因となる疾患の原因遺伝子を探索するには、ゲノムの情報は有用である。図３（Ｅ）は遺伝子等同士が階層構造にある場合で、階層構造の例として図７のオントロジー、図８の酵素（ＥＣ：Ｅｎｚｙｍｅ Ｃｏｍｍｉｓｓｉｏｎ）、図９のスーパーファミリーの関係がある。図７はＡＤＰＲＴと呼ばれる遺伝子のオントロジーである。オントロジーとは遺伝子配列や蛋白質配列解析等に基づき、遺伝子の機能を定義した辞書である。オントロジーの詳細については、Ｔｈｅ ｇｅｎｅ ｏｎｔｏｌｏｇｙ ｃｏｎｓｏｒｔｉｕｍ、Ｇｅｎｅ ｏｎｔｏｌｏｇｙ

：ｔｏｏｌ ｆｏｒ ｔｈｅ ｕｎｉｆｉｃａｔｉｏｎ ｏｆ ｂｉｏｌｏｇｙ．Ｎａｔｕｒｅ Ｇｅｎｅｔｉｃｓ ２５，ｐｐ．２５－２９ （２０００） を参照。遺伝子オントロジーによるとＡＤＰＲＴ遺伝子は、ＤＮＡ修復（ＤＮＡ ｒｅｐａｉｒ）、ＡＤＰ－リボシル化（ＡＤＰ－ｒｉｂｏｓｙｌａｔｉｏｎ）の機能を有することが分かる。ＤＮＡ修復はＤＮＡ代謝（ＤＮＡ ｍｅｔａｂｏｌｉｓｍ）の一つであり、ＤＮＡ代謝は核酸等代謝（ｎｕｃｌｅｏｂａｓｅ，ｎｕｃｌｅｏｓｉｄｅ，ｎｕｃｌｅｏｔｉｄｅ ａｎｄ ｎｕｃｌｅｉｃ ａｃｉｄ ｍｅｔａｂｏｌｉｓｍ）の一つである。

【００１８】

図７のオントロジー右端括弧中の数字は登録遺伝子数を示す。３８６個の遺伝子がＤＮＡ修復機能を有する遺伝子として現在登録されており、ＤＮＡ修復を含むＤＮＡ代謝には１１３８個の遺伝子が登録されている。遺伝子オントロジーでは汎用的な大分類から詳細な小分類へと階層構造を形成していることが、登録遺伝子数からも見て取れる。図８に階層構造の二つ目の例である酵素の例を示す。酵素はその構造や機能に基づいたＥＣ番号（Ｅｎｚｙｍｅ Ｃｏｍｍｉｓｓｉｏｎ）によって分類されている。ＥＣ１からＥＣ６まであり、ＥＣ１はオキシドレダクターゼ（Ｏｘｉｄｏｒｅｄｕｃｔａｓｅｓ）、ＥＣ２はトランスフェラーゼ（Ｔｒａｎｓｆｅｒａｓｅｓ）、ＥＣ３はハイドロラーゼ（Ｈｙｄｒｏｌａｓｅｓ）、ＥＣ４はリアーゼ（Ｌｙａｓｅｓ）、ＥＣ５はイソメラーゼ（Ｉｓｏｍｅｒａｓｅｓ）、ＥＣ６はリガーゼ（Ｌｉｇａｓｅｓ）である。ＥＣ１から６は更に階層構造に従った分類がなされている。図８ではＥＣ６の例をしめすが、ＥＣ６はＥＣ６．１からＥＣ６．５に分類される。ＥＣ６．３は更にＥＣ６．３．１からＥＣ６．３．５までに分類される。ＥＣ６．３．３の場合、実際の酵素はＥＣ６．３．３．１からＥＣ６．３．３．３である。

【００１９】

図９に階層構造の三つ目の例であるスーパーファミリーの例を示す。スーパーファミリーとは塩基配列の解析から得られるモチーフ、ドメイン構造から、類似の蛋白質立体構造や機能を有することが予想される一連の遺伝子群のことである。なおモチーフとは構造やパターンの要素のことであるが、ここでは各種のタンパク質のアミノ酸配列中に認められる一定の構造を指す。モチーフは互いに機能が異なる幅広いタンパク質に共通して見られる構造である。なおタンパク質にはドメインと呼ばれる構造があるが、タンパク質のドメインはモチーフの種々の組み合わせでできている。一般にモチーフは、タンパク質のドメインよりは小さい構造単位である。モチーフには、例えばヘリックスーターンーヘリックスやジンクフィンガーと呼ばれるＤＮＡ結合構造モチーフなどがある。図９は薬物代謝酵素ＣＹＰ遺伝子群の例を示している。ＣＹＰ遺伝子群は全体で約５０個ほど存在するが、それらの遺伝子群は、構造や機能ごとにＣＹＰ１Ａ１，ＣＹＰ１Ａ２などのグループに分類されている。

【００２０】

図３（Ｆ）は遺伝子等同士がネットワーク関係にある場合で、ネットワークの例として図１０の文献情報、図１１の蛋白質相互作用、図１２の代謝経路の関係がある。図１０にネットワークの一つ目の例である文献情報に基づく相互関係について示す。この方法は、二つの遺伝子名が同一の文献データベースの同一文章内に存在する数が多いほど両者の相互関係が強いとするものであり、例えばＰｕｂＧｅｎｅデータベースを用いて相互関係スコアを得ることができる。Ｔｏｒ－Ｋｒｉｓｔｉａｎ Ｊｅｎｓｓｅｎら、Ａ ｌｉｔｅｒａｔｕｒｅ ｎｅｔｗｏｒｋ ｏｆ ｈｕｍａｎ ｇｅｎｅｓ ｆｏｒ ｈｉｇｈ－ｔｈｒｏｕｇｈｐｕｔ ａｎａｌｙｓｉｓ ｏｆ ｇｅｎｅ ｅｘｐｒｅｓｓｉｏｎ，Ｎａｔｕｒｅ Ｇｅｎｅｔｉｃｓ，ｖｏｌ．２８，ｐｐ２１－２８参照。図１０の各丸印は遺伝子を、丸印同士をつなぐ線は、相互関係があることを示している。線に並んで記載された数字は相互関係スコアを表す。図１０では相互関係スコアとは、線でつながれた二つの遺伝子が医学文献データベースＭＥＤＬＩＮＥの同一アブストラクト文中に存在した件数を示している。図１０の中央にある遺伝子ＡＤＰＲＴと相互関係が強いものは、ＴＰ５３，ＣＦＴＲ，ＥＥＦ２，ＦＲＡ１Ｈ，ＳＰ１，ＤＡＦの６遺伝子であり、相互関係ス

コアは６遺伝子とも１である。このＰｕｂＧｅｎｅでは文献データベースとして米国ＮＣＢＩのＭＥＤＬＩＮＥやＯＭＩＭを用いているが、その他の文献データベースでもかまわない。図１１にネットワークの２つ目の例である蛋白質相互作用をしめす。米国カリフォルニア大学ロサンゼルス校のＤＩＰ（Ｄａｔａｂａｓｅ ｏｆ Ｉｎｔｅｒａｃｔｉｎｇ Ｐｒｏｔｅｉｎｓ）などのような、タンパク質相互作用データベースを用いて蛋白質相互作用を調べることができる。ＤＩＰについてはＩ．Ｘｅｎａｒｉｏｓら、ＤＩＰ：ｔｈｅ ｄａｔａｂａｓｅ ｏｆ ｉｎｔｅｒａｃｔｉｎｇ ｐｒｏｔｅｉｎｓ，Ｎｕｃｌｅｉｃ Ａｃｉｄ Ｒｅｓｅａｒｃｈ，ｖｏｌ．２８，２８９－２９１，２００参照。タンパク質相互作用データベースにおいても、相互作用するタンパク質同士が線で結合されている。相互作用の強さは、例えば解離定数（Ｄｅｓｓｏｃｉａｔｉｏｎ Ｃｏｎｓｔａｎｔ）が実験的に求められていれば、分子同士の結合力が分かるので、結合力の強い方をより相互作用が大きいと見なす。また１回の実験で確認された相互作用より、２回以上の複数回の実験で確認された相互作用を、より相互作用が強いと見なしてもよい。なおＤＩＰ以外のタンパク質相互作用データベースを用いてもかまわない。図１１にネットワークの３つ目の例である代謝経路をしめす。代謝経路の詳細はＫＥＧＧ（ｈｔｔｐ：／／ｗｗｗ．ｋｅｇｇ．ｋｙｏｔｏ－ｕ．ａｄ．ｊｐ）などの代謝経路データベースを参照。図３（Ｃ）のパスウェイとこの代謝経路との違いは、補酵素（ｃｏ－ｅｎｚｙｍｅ）と呼ばれる反応を媒介する複数の蛋白質が代謝反応では関与するため、単純に上流から下流という一方向の関係にとどまらない複雑な構造をとる点である。

【００２１】

図３から図１２はゲノム、トランスクリプトーム、プロテオームといった遺伝子情報における相互関係を示している。しかし遺伝子情報は生物情報の一部である。図１３に示すように、生物情報には、酵素間相互作用（酵素全体像：エンザイモーム：Ｅｎｚｙｍｏｍｅ，代謝全体像：メタボローム：Ｍｅｔａｂｏｌｏｍｅ）、相互作用（相互作用全体像：インタラクトーム：Ｉｎｔｅｒａｃｔｏｍｅ）、時間的空間的局在（局在全体像：ローカリゾーム：Ｌｏｃａｌｉｚｏｍｅ）、表現型（表現型全体像：フェノーム：Ｐｈｅｎｏｍｅ）にいたる階層構造がある。Ｍ．Ｖｉｄａｌ、Ａ ｂｉｏｌｏｇｉｃａｌ ａｔｌａｓ ｏｆ ｆｕｎｃｔｉｏｎａｌ ｍａｐｓ．Ｃｅｌｌ １０４，ｐｐ．３３３－３３９（２００１）を参照。本願明細書の解析法は、ゲノム、トランスクリプトーム、プロテオーム内の相互作用に留まらず、図１３に示す生物情報全般における相互作用についても、適用することができる。

【００２２】

図１４と図１５に本発明を実現するためのコンピュータ化された方法を示す模式図を示す。図１４は、クロスバリデーション値と既定値とを比較することで繰り返し回数を決定する、本発明を実現するためのコンピュータ化された方法を示す模式図である。

【００２３】

図１５は、全てのクロスバリデーション値を保存して比較する、本発明を実現するためのコンピュータ化された方法を示す模式図である。図１４と図１５において、遺伝子発現データを受け取る工程、クラス情報を受け取る工程、データマイニング手法を用いてクラス分類に関連する遺伝子群を抽出する工程、前記遺伝子群に注釈付け（アノテーション）を行う工程、遺伝子アノテーションに基づき前記クラス分類に関連する遺伝子群の共通規則を抽出する工程、前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程は共通である。図１４と図１５では遺伝子発現データを受け取った後に、正規化とフィルタリング処理を行ってもよいし行わなくてもよいが、特にフィルタリングはＤａｔａ Ｓｅｌｅｃｔｉｏｎ、Ｄａｔａ Ｃｌｅａｎｓｉｎｇというデータマイニング結果を左右する工程なので行うことが望ましい。

【００２４】

ＤＮＡチップのデータ解析において有用なフィルタリングの例を表１に示す。フィルタリング法の１例であるしきい値法は、前述のように信号強度が小さい（例えば５００以下の）場合、しきい値イコール５００とし、５００以下の信号強度を５００にするかもしくは

ゼロにする操作のことである。しきい値の決定には経験に基づく方法、ノンパラメトリック検定、例えばＷｉｌｃｏｘｏｎ符号付順位検定に基づく方法などがある。

【００２５】

【表１】

表１

フィルタリングの例

(1)しきい値法 (Threshold method)
(2)相関係数法 (Correlation method)
(3)分散分析法 (Modified ANOVA)
(4)ベイズ推定法 (Bayesian T-test)

フィルタリング法の１例である相関係数法は、まず（１）実データの全遺伝子に対し、複数サンプルから２個のサンプルを無作為に抽出して、相関係数を計算する。（２）続いて人工的に作成したランダムデータに対し、前記と同様に相関係数を計算する。相関係数にはピアソンの相関係数を用いる。（３）ランダムデータでの相関係数の確率分布と、実データでの相関係数との確率分布を比較する。（４）ランダムデータの分布から見て有意に大きい相関もしくは逆相関を有する遺伝子群のみをデータマイニングの入力として用いる方法である。

【００２６】

フィルタリング法の１例である分散分析法は標準的なＡＮＯＶＡ（Ａｎａｌｙｓｉｓ ｏｆ Ｖａｒｉａｎｃｅ）と似ている。但し標準的なＡＮＯＶＡは、データ同士が互いに独立であることを仮定して、Ｆ分布を判定に用いる。実際の遺伝子では、遺伝子データ同士が何らかの相関があることが十分予想されるので、通常のＡＮＯＶＡを用いることは誤りである。そこで、Ｆ分布に相当する分布を実データからブートストラップ法などを用いて模擬的に作成する方法が取られることが多い。

【００２７】

フィルタリング法の１例であるベイズ推定法について説明する。以下Ｃｙ３とＣｙ５という二色の蛍光色素で、２種類のサンプルを染色して同時にハイブリダイゼーションするＤＮＡチップ実験を想定する。予め、同一ＲＮＡを二つに分けて、片方をＣｙ３、もう片方をＣｙ５で標識し、同一チップ上で競合ハイブリさせる実験（チップ枚数は５、１０枚程度必要）を行うことを考える。この実験のデータは、Ｃｙ５／Ｃｙ３の平均値＝ｃは１．０であるが、分散σ^2については未知である正規母集団Ｎ（ｃ、σ^2）から大きさｎの無作為標本｛Ｙ１，Ｙ２，・・・，Ｙｉ，・・・，Ｙｎ｝を抽出し、その観測値ｙ＝（ｙ１，ｙ２，・・・，ｙｉ，・・・，ｙｎ）を得たことに相当する。チップ間差、色素間差、ハイブリダイゼーションの個人差等の誤差要因は、実験間で相関がないと考え

られるため｛Ｙ１，Ｙ２，・・・，Ｙｉ，・・・，Ｙｎ｝は相互に独立（ｉｎｄｅｐｅｎｄｅｎｔｌｙ　ａｎｄ　ｉｄｅｎｔｉｃａｌｌｙ　ｄｉｓｔｒｉｕｔｅｄ；　ｉ．ｉ．ｄ）と言える。

【００２８】

【式１】

$$Yi \sim i.i.d.\ N(c,\ \sigma^2) = N(1,\ \sigma^2) \qquad (式1)$$

本ベイズ推定の最終目的は、Ｙｉ　のσ^2を推定することにある。本推定により母集団Ｎ（１、σ^2）のσ^2を推定すれば、式１より、Ｙｉのσ^2も同一である。ベイズ推定ではσ^2は分布を持つとするので、あくまでσ^2の推定値が得られる。推定値としては、平均、モード（度数が最大となる点）や、信頼区間として最高密度信頼区間（Ｈｉｇｈｅｓｔ　Ｄｅｎｓｉｔｙ　Ｒｅｇｉｏｎ；ＨＤＲ）が得られる。９０％最高密度区間は、いわば未知母数σ２の９０％区間のうちで最も短く、また事後分布のピーク値（事後モード）を必ず含み、かつ区間の両端における事後密度が等しくなるものである。

式１が成り立つとき、｛Ｙ１，Ｙ２，・・・，Ｙｉ，・・・，Ｙｎ｝の同時確率密度分布（ａ１　＜　Ｙ１　≦　ｂ１，　ａ１　＜　Ｙ１　≦　ｂ１，・・・、ａｎ　＜　Ｙｎ　≦　ｂｎが同時に満たされる確率密度分布）ｐ（ｙ‘｜ｃ，　σ^2）は、

【００２９】

【式２】

$$p(y' \mid c,\ \sigma^2) = \prod_{i=1}^{n} (1/\sqrt{2\pi}\sigma) \exp(-1/2\sigma^2 \times (yi-c)^2) \qquad (式2)$$

のように正規分布を掛け合わせた形式で書ける。但しΠは積記号である。従って、観測値ベクトルｙ＝（ｙ１，　ｙ２，　・・・，ｙｉ，・・・，ｙｎ）が与えられたときの尤度関数ｌ（σ^2｜　ｙ）は、

【００３０】

【式３】

$$l(\sigma^2 \mid y) \propto \prod_{i=1}^{n} (1/\sigma) \exp(-1/2\sigma^2 \times (y-c)^2) \qquad (式3)$$

【式４】

$$l(\sigma^2 \mid y) \propto (\sigma^2)^{-n/2} \times \exp(-ns^2/2\sigma^2) \qquad (式4)$$

となる。ただし、ｓ^{2}は母平均ｃを中心とした観測値分布であり、下記の式で表される。

【００３１】

【式５】

$$s^2 = (1/n) \times \sum_{i=1}^{n} (yi-c)^2 \qquad (式5)$$

ここで、分散の事前分布p（σ^2）として、無情報事前分布（ｎｏｎｉｎｆｏｒｍａｔｉｖｅ　ｐｒｉｏｒ　ｄｉｓｔｒｉｂｕｔｉｏｎ）を仮定する。この仮定は母数に関して無知であるとすることで、事前分布についての恣意性を出来る限り排除し、事後分布はできるだけデータによって支配されるようにする点で妥当である。無情報事前分布として、局所一様事前分布を用いるのが一般的である。局所一様分布とは、未知母数を二乗しようが、三乗しようが、対数値をとろうが、事前情報の漠然性を表すために少なくとも局所的には一様に分布するような分布のことである。具体的にはフィッシャー情報量の平方根に比例するように定めればよいことが分かっている。事前分布として局所一様分布を用いるならば、

【００３２】

【式６】

$$p(\sigma^2) \propto \sigma^{-2} \qquad (式6)$$

とすれば良い。つまりσ^2の事前分布p（σ^2）は、σ^{-2}すなわち定数とする。次に事後分布を考える。ベイズの定理より

【００３３】

【式７】

$$p(\sigma^2 \mid y) \propto l\ (\sigma^2 \mid y)\ p(\sigma^2) \qquad (式7)$$

が成立するので、

【００３４】

【式８】

$$p(\sigma^2 \mid y) \propto (\sigma^2)^{-(n/2+1)} \times \exp(-ns^2/2\sigma^2) \qquad (式8)$$

より、σ^2の事後分布p（σ^2 | y）は、χ^{-2}（ｎ、ｎｓ^2）と等しい分布になる。なおχ^{-2}（ν、λ）とは、尺度母数λをもつ自由度νの逆カイ二乗分布と呼ばれる分布である。χ^{-2}（ν、λ）の平均はλ／（ν－２）、モード（度数が最大となる点）はλ／（ν＋２）となることが分かっているので、σ^2の点推定値として、

【００３５】

【式９】

平均値を基準とした場合：$\sigma^2 = ns^2$／（ｎ－２）　　（式９）

【式１０】

モードを基準した場合：$\sigma^2 = ns^2 / (n+2)$　　（式１０）

を考えることができる。また式８より事後的に、

【００３６】

【式１１】

$ns^2 / \sigma^2 \sim \chi^2(n)$　　（式１１）

の関係が得られる。式１１のχ^2（ｎ）とは、自由度ｎのカイ二乗分布である。ここでは、ｎｓ^{2}が固定値（観測値）、σ^2が確率変数になっている。式１１と数表を用いることで、ＨＤＲを求めることができる。式９，１０，１１をもちいることで、２種類のサンプル間の遺伝子発現比率データが得られたとき、その比率が１．０と比較してどの程度、統計的有意に異なるかを知ることができる。１．０より有意に異なる比率を有する遺伝子群のみをデータマイニングの入力として使用すれば良い。例えば５回の実験で、$y_1 = 1.4$、$y_2 = 0.89$、$y_3 = 1.24$、$y_4 = 0.91$、$y_5 = 1.04$が得られたとすると、$s^2 = 0.04788$である。平均値を基準とした点推定（式９）より、Ｙｉ～Ｎ（１，　０．０７９８）となる。またモードを基準とした点推定より（式１０）より、Ｙｉ～Ｎ（１，　０．０３４２）となる。式１１より、$\sigma^2 \sim 0.2394\chi^{-2}$（５）、となることから、数表を用いることで$\sigma^2$の９０％ＨＤＲは０．０１９－０．１７７となる。なお式１から式１１までに示したベイズ推定法は統計手法の一つの方法であるので、ベイズ推定法以外の方法、例えば、ネイマン・ピアソン流の推定法を用いて同様の推定を行ってもよい。

【００３７】

発現データを読み込んだ後、正規化、フィルタリングを行う、またクラス情報を読み取ることで、データマイニングに入力するためのデータ形式である発現マトリクスを作成することができる。発現マトリクス形式を図１６に示す。遺伝子アノテーションを行（Ｒａｗ）、サンプルアノテーションを列（Ｃｏｌｕｍｎ）、対応する発現レベル（例えば前述のＣｙ５とＣｙ３の比率：Ｃｙ５／Ｃｙ３）をマトリクスにした構造である。この発現マトリクス形式のサンプルアノテーション部位にクラス情報を付加することで、データマイニングに適した構造となる。この発現マトリクスは例えばＣＳＶ形式やタブ区切り形式などの形式でファイルに保存することができる。

【００３８】

次のデータマイニング工程は、１回目であれば第１次データマイニング、以後マイニングを反復して行った（イタレーションした）場合その回数がｎ回目であればｎ次データマイニング工程と呼ばれる。チップデータに適したマイニング方法を４つ表２に示した。いずれも教師付学習法（スーパーバイズドメソッド）において代表的な方法である。最短近傍法、判別分析についてはＳｐｌｕｓ等の統計計算パッケージ（Ｓｐｌｕｓのｃｌａｓｓパッケージ）を用いて実行することができる。またサポートベクターマシンについてもフリーソフトでＳｐｌｕｓのクローンであるＲ言語のｅ１０７１パッケージを用いて実行することができる。次に特徴ルール法について説明する。

【００３９】

【表２】

表2

マイニングの例

教師付学習法(Supervised Methods)

(1)最短近傍法 (Nearest Neighbor)
(2)判別分析 (Discriminant Analysis)
(3)サポートベクターマシン (Support vector machine)
(4)特徴ルール法 (Characteristic Rule)

特徴ルール法は特開平８－７７０１０に、データ分析方法として開示されている。特徴ルール法では、複数の属性項目からなるサンプルの集合を対象データとする。全てのサンプルは互いに同一の属性項目を持つ。それぞれの属性項目が取り得る属性値はサンプル数と比較して少数の離散値であることが要求される。典型的には、３通り程度の記号値である。元の分析対象データが実数値データである場合、適当な境界で値の範囲を区切り、「大」「中」「小」といった記号値に置き換える等の方法で離散化する。

【００４０】

特徴ルール法を実施する際には、属性項目の１個を選び、「結論項目」とする。また、結論項目の取り得る属性値のうちの１個を選び、「結論項目値」とする。また、その他の複数の属性項目を選び、「条件項目」とする。

【００４１】

特徴ルール法では、「ＩＦ（条件部）ＴＨＥＮ（結論部）」という形式のＩＦ－ＴＨＥＮルールを生成する。ルールの条件部は、条件項目とその属性値の組、すなわち述語であり、複数の述語が同時に条件部に現れることを許すが、典型的には３個程度に制限する。また、結論部は、結論項目と結論項目値からなる述語である。これにより、結論部の述語はただ１つに決定され、一方、条件部は様々な述語の組み合わせを取り得る。したがって、生成し得るＩＦ－ＴＨＥＮルールの数は、一般に多数になる。これら多数のＩＦ－ＴＨＥＮルールの中から、対象データの特徴をよく表している比較的少数のルールを探索することが特徴ルール法の目的である。各ＩＦ－ＴＨＥＮルールが対象データの特徴をどの程度よく表しているかを評価するために、以下の評価尺度を用いる。条件部をＡ（複数の述語の組み合わせを含む）、結論部をＢとする、「ＩＦ　Ａ　ＴＨＥＮ　Ｂ」というルールの評価尺度μ（Ａ→Ｂ）を次式のように定義する。

μ（Ａ→Ｂ）＝Ｐ（Ａ）＾β　×　ｌｏｇ［Ｐ（Ｂ｜Ａ）／Ｐ（Ｂ）］

ここで、Ｐ（Ａ）＾β　は　Ｐ（Ａ）のβ乗を意味する。Ｐ（Ａ）は対象データの中で条件部Ａが満足される確率、すなわち、対象データ全体の中で、Ａという条件を満たすサン

ブルの割合を表す。同様に、Ｐ（Ｂ）は結論部Ｂが満足される確率、Ｐ（Ｂ｜Ａ）は、Ａを満たすという条件の下で結論部Ｂが満足される条件付確率を表す。βは使用者が指定するパラメータで、０以上、１以下の実数値である。この評価尺度によって与えられる評価値が大きいルールほど、対象データの特徴をよく表していると見なす。また、上記の評価尺度の定義式におけるＰ（Ａ）をカバー率、Ｐ（Ｂ｜Ａ）をヒット率と呼び、これらは、取り出されたルールを使用者が解釈する際の手がかりとして用いられることがある。

【００４２】

生成され得る多数のＩＦ－ＴＨＥＮルールの中から、評価値の大きい、比較的少数のルールを取り出すアルゴリズムはいくつか考えられるが、「総当たり法」は、もっとも単純な方法の１つである。これは、取り出すルール数（例えば、１０）をあらかじめ定めておき、そして、条件部に同時に現れる述語数の上限（例えば、３）を定め、その範囲内で可能な全てのＩＦ－ＴＨＥＮルールを生成、評価し、その中で評価値の大きい上位のルールを、あらかじめ定めた数だけ取り出すというものである。

【００４３】

特徴ルール法を、クラス分類に関する遺伝子群の抽出に用いるには、対象データにおいて、クラス分類に関する属性をルールの結論項目とし、遺伝子アノテーションに該当する属性を条件項目とする。これにより、クラス分類に関して重要な遺伝子を条件部に持つＩＦ－ＴＨＥＮルールを得る。

【００４４】

データマイニングの結果、クラス分類において重要な（クラス分類を行う規則に関連する）遺伝子群を抽出するが、一般に正解率が高い順（エラー率の低い順）から、１つ以上で１０個から２００個、好ましくは２０個から５０個の遺伝子を選択する。この重要遺伝子群抽出には前述の特徴ルール法が最も優れている。

【００４５】

特徴ルール法以外の例として、例えばＢＳＳ／ＷＳＳ比を用いる方法がある。なおＢＳＳはＢｅｔｗｅｅｎ－ｇｒｏｕｐ　ｓｕｍ　ｏｆ　ｓｑｕａｒｅｓ、ＷＳＳはＷｉｔｈｉｎ－ｇｒｏｕｐ　ｓｕｍ　ｏｆ　ｓｑｕａｒｅｓの略であり、遺伝子ｊのＢＳＳ／ＷＳＳ比は下記の式で定義される。

【００４６】

【式１２】

$$BSS(j)/WSS(j) = \sum_i \sum_k I(yi = k)(m(xkj) - m(x.j))^2 \Big/ \sum_i \sum_k I(yi = k)(xij - m(xkj))^2 \qquad (式１２)$$

但し、ｍ（ｘ．ｊ）は全検体における遺伝子ｊの平均発現量、ｍ（ｘｋｊ）は、クラスｋに属する検体における遺伝子ｊの平均発現量、ｘｉｊは遺伝子発現ﾃﾞｰﾀマトリクス、Ｉ（ｙｉ　＝　ｋ）は、ｙｉ　＝　ｋのとき１、それ以外は０となる関数である。式１２で表されるＢＳＳ／ＷＳＳ比が大きいほど、クラス内部の相違と比較してクラス間の相違が大きいので、クラシフィケーションの際にはＢＳＳ／ＷＳＳ比が大きい遺伝子、例えば上位２０から５０個を用いることが望ましい。

【００４７】

次にデータマイニングにより抽出されたクラス分類において重要な遺伝子群についてアノテーション付けを行う。アノテーションには図３から図１３までの様々な相互作用のうちいずれか１つ以上を行う。表３には例として遺伝子オントロジーを用いてアノテーションを行った例を示す。データマイニングの結果抽出された遺伝子群がＰｒｏｂｅ１からＰｒｏｂｅ５であった場合、もともとＤＮＡチップ設計時に既知である遺伝子クラスター番号（Ｕｎｉｇｅｎｅ番号）、塩基配列番号（Ｇｅｎｂａｎｋ番号）、遺伝子名に加え、遺伝

子アノテーション工程において、染色体番号や遺伝子オントロジーを公共データベース等から検索し、表３を作成することができる。

【００４８】

【表３】

表3

アノテーションの例

プローブ番号	Unigene番号	Genbank番号	遺伝子名	染色体番号	遺伝子オントロジー
Probe1	Mm.19904	AA413214	Bcl2-associated X protein	7	integral membrane protein apoptosis regulator apoptosis
Probe2	Mm.98	AA450909	proteasome (prosome, macropain) subunit, beta type 6	11	20S core proteasome peptidase, proteasome endopeptidase ubiquitin-dependent protein degradation
Probe3	Mm.5341	AA065510	defensin beta 1	8	extracellular antimicrobial peptide defense response, xenobiotic metabolism
Probe4	Mm.24816	AI643210	coagulation factor II (thrombin) receptor	13	membrane, integral membrane protein blood coagulation factor G-protein coupled receptor protein signaling pathway
Probe5	Mm.850	AA106360	signal recognition particle 14 kDa (homologous Alu RNA-binding protein)	2	signal recognition particle RNA binding protein targeting

続いて遺伝子アノテーションから共通性や規則性を抽出する工程を行う。表３の遺伝子群はあるクラス分類において重要であることがデータマイニングにより分かった遺伝子群であるとする。この表３の遺伝子群から共通した性質や特徴を抽出することが、遺伝子アノテーションから共通規則を抽出する工程の目的である。例えば、表３の遺伝子オントロジーに対して、複数の遺伝子において共通に見られるオントロジーがないかを検索する。するとＰｒｏｂｅ１とＰｒｏｂｅ４で、ｉｎｔｅｇｒａｌ ｍｅｍｂｒａｎｅ ｐｒｏｔｅｉｎが共通して見られる。そこで表３からは、蛋白質として膜（ｉｎｔｅｇｒａｌ ｍｅｍｂｒａｎｅ）に存在する遺伝子群がクラス分類に関わっているという規則が潜んでいる可能性を見出すことができる。次のデータマイニングでは、例えば、ＤＮＡチップ上に搭載されているプローブに対し、ｉｎｔｅｇｒａｌ ｍｅｍｂｒａｎｅ ｐｒｏｔｅｉｎ（ＧＯ番号：００１６０２１）に対応するプローブという拘束条件を設けて、それに該当する遺伝子群のみでデータマイニング・クロスバリデーションすることで、ｉｎｔｅｇｒａｌ ｍｅｍｂｒａｎｅ ｐｒｏｔｅｉｎがどの程度、クラス分類に寄与するかを定量的に把握することができる。またｉｎｔｅｇｒａｌ ｍｅｍｂｒａｎｅ ｐｒｏｔｅｉｎ（Ｇ

O番号：００１６０２１）の上階層は、ｍｅｍｂｒａｎｅ　（ＧＯ番号：００１６０２０）である。そこで別のデータマイニングでは、ＤＮＡチップ上に搭載されているプローブに対し、ｉｎｔｅｇｒａｌ　ｍｅｍｂｒａｎｅ　ｐｒｏｔｅｉｎ（ＧＯ番号：００１６０２１）もしくはｍｅｍｂｒａｎｅ　（ＧＯ番号：００１６０２０）に対応するプローブという拘束条件を設けて、それに該当する遺伝子群のみでデータマイニング・クロスバリデーションしてもよい。仮に“ｉｎｔｅｇｒａｌ　ｍｅｍｂｒａｎｅ　ｐｒｏｔｅｉｎを含む”を拘束条件としてマイニングを行った場合と“ｍｅｍｂｒａｎｅを含む”を拘束条件としてマイニングを行った場合とでクロスバリデーション結果（正解率もしくはエラー率）を比較した場合に、前者の正解率が後者より良ければクラス分類には、ｍｅｍｂｒａｎｅでなく、ｉｎｔｅｇｒａｌ　ｍｅｍｂｒａｎｅ　ｐｒｏｔｅｉｎが重要であることが分かる。このように遺伝子アノテーションから共通規則を抽出する工程とは、前記のオントロジーのように、図３から図１３に示した相互関係を、表３のようなクラス分類の際に重要な遺伝子リストから検索する工程である。規則性抽出において、表４から表６に示したルールに従うことでより一般性の高い規則を見出すことができる。この見出された規則を拘束条件として、この拘束条件のもとにデータマイニングを行う工程が、次の、「前記共通規則に基づく拘束条件をもちいたデータマイニングを行う工程」である。この工程では、ｎ次マイニングより得られた重要遺伝子群とそのアノテーションを用いて（ｎ＋１）次マイニングを行う。このようにマイニングを繰り返す意義は、ルール生成で得られた遺伝子群に、アノテーションによって情報を付加し、これらの情報を各遺伝子の属性と考え、共通する特徴をさらに解析する。更にこの解析サイクルを繰り返すことで共通する特徴を徹底して探索することができることである。なお、規則の発明に際し、リガンドーレセプタ関係についてはＴ．Ｇ．Ｇｒａｅｂｅｒ　ａｎｄ　Ｄ．　Ｅｉｓｅｎｂｅｒｇ、Ｂｉｏｉｎｆｏｒｍａｔｉｃ　ｉｄｅｎｔｉｆｉｃａｔｉｏｎ　ｏｆ　ｐｏｔｅｎｔｉａｌ　ａｕｔｏｃｒｉｎｅ　ｓｉｇｎａｌｉｎｇ　ｌｏｏｐｓ　ｉｎ　ｃａｎｃｅｒｓ　ｆｒｏｍ　ｇｅｎｅ　ｅｘｐｒｅｓｓｉｏｎ　ｐｒｏｆｉｌｅｓ．　Ｎａｔｕｒｅ　Ｇｅｎｅｔｉｃｓ　２９，　ｐｐ．２９５－３００　（２００１）　を、蛋白質相互作用関係についてはＨ．　Ｇｅら、Ｃｏｒｒｅｌａｔｉｏｎ　ｂｅｔｗｅｅｎ　ｔｒａｎｓｃｒｉｐｔｏｍｅ　ａｎｄ　ｉｎｔｅｒａｃｔｏｍｅ　ｍａｐｐｉｎｇ　ｄａｔａ　ｆｒｏｍＳａｃｃｈａｒｏｍｙｃｅｓ　ｃｅｒｅｖｉｓｉａｅ．　Ｎａｔｕｒｅ　Ｇｅｎｅｔｉｃｓ　２９，　ｐｐ．４８２－４８６　（２００１）をそれぞれ参考にした。但し、前述の２件の公知例はどちらも、最初の１回目のマイニングに際し入力する発現マトリクスを改変する方法を提示しているのみである。本願明細書に記載されたように前回のマイニング結果を元に、更にマイニングを連続的に反復実行することはない。これでは一般性・頑強性（ロバストネス）に優れた知識獲得を行うことは困難である。本発明は、一般性・頑強性に優れた知識獲得を自動で行える点で公知技術とは根本的に異なっている。

【００４９】

【表４】

表4

n次マイニングの結果を用いて(n+1)次マイニングを行う際のルール(その1)

相互関係の種類	一次マイニングで発見された関連候補遺伝子	二次マイニングにおける探索シード
二項関係	リガンド遺伝子	○レセプター遺伝子 ○リガンドレセプター対
	レセプター遺伝子	○リガンド遺伝子 ○リガンドレセプター対
パスウェイ	パスウェイP_A上の遺伝子	○パスウェイP_Aの上流遺伝子 ○パスウェイP_Aの下流遺伝子 ○相関パスウェイP_B上の遺伝子 ○上記いずれかと一次マイニング遺伝子とのペア
ゲノム	染色体C_A上の遺伝子	○染色体C_A上の隣接遺伝子 ○上記と一次マイニング遺伝子とのペア
階層構造	オントロジーがO_Aの遺伝子	○ O_Aの上階層のオントロジーを有する遺伝子 ○上記と一次マイニング遺伝子とのペア

【表5】

表5

n次マイニングの結果を用いて(n+1)次マイニングを行う際のルール(その2)

相互関係の種類	一次マイニングで発見された関連候補遺伝子	二次マイニングにおける探索シード
階層構造	酵素EC_Aの遺伝子	○酵素EC_Aの上階層の遺伝子 ○酵素EC_Aと同一グループに属する遺伝子 ○上記いずれかと一次マイニング遺伝子とのペア
	スーパーファミリーSF_Aに属する遺伝子	○同一スーパーファミリーSF_Aに属する遺伝子(モチーフ、ドメイン構造が相互に類似) ○上記と一次マイニング遺伝子とのペア
ネットワーク	遺伝子A	○文献情報から遺伝子Aとの関連が予想される遺伝子B ○上記と一次マイニング遺伝子(遺伝子A)とのペア

【表6】

表6

n次マイニングの結果を用いて(n+1)次マイニングを行う際のルール(その3)

相互関係の種類	一次マイニングで発見された関連候補遺伝子	二次マイニングにおける探索シード
ネットワーク	遺伝子A	○蛋白質相互作用データから遺伝子Aとの関連が予想される遺伝子B ○上記と一次マイニング遺伝子(遺伝子A)とのペア
	遺伝子A	○代謝経路情報から遺伝子Aとの関連が予想される遺伝子B ○上記と一次マイニング遺伝子(遺伝子A)とのペア

表４から表６について説明する。表４から表６において遺伝子間の相互関係が二項関係でありかつリガンドーレセプターの関係であった場合、クラス分類関連抽出遺伝子がリガンド遺伝子であれば、アノテーションから抽出された共通性や規則性に基づく拘束条件をレセプター遺伝子もしくはリガンド遺伝子とレセプター遺伝子の組を含む条件とし、クラス分類関連抽出遺伝子がレセプター遺伝子であれば、前記拘束条件をリガンド遺伝子もしくはリガンド遺伝子とレセプター遺伝子の組を含む条件とする。

【００５０】

表４から表６において遺伝子間の相互関係がパスウェイであった場合、クラス分類関連抽出遺伝子がパスウェイＰＡ上にあれば、アノテーションから抽出された共通性や規則性に基づく拘束条件をパスウェイＰＡ上の上流遺伝子、下流遺伝子、パスウェイＰＡと相関するパスウェイＰＢ上の遺伝子、もしくは前記いずれかの遺伝子の組を含む条件とする。

【００５１】

表４から表６において遺伝子間の相互関係がゲノムであった場合、クラス分類関連抽出遺伝子が染色体ＣＡ上にあれば、アノテーションから抽出された共通性や規則性に基づく拘束条件を染色体ＣＡ上の隣接遺伝子、もしくは前記の抽出された遺伝子と前記の隣接遺伝子との組を含む条件とする。

【００５２】

表４から表６において遺伝子間の相互関係が階層構造でかつオントロジーであった場合、クラス分類関連抽出遺伝子のオントロジーがＯＡであれば、アノテーションから抽出された共通性や規則性に基づく拘束条件をオントロジーＯＡの上階層のオントロジーを有する遺伝子、もしくは前記の抽出された遺伝子と前記の上階層のオントロジーを有する遺伝子との組を含む条件とする。

【００５３】

表４から表６において遺伝子間の相互関係が階層構造でかつ酵素（ＥＣ）であった場合、クラス分類関連抽出遺伝子のＥＣ番号がＥＣＡであれば、アノテーションから抽出された

共通性や規則性に基づく拘束条件を酵素ＥＣＡの上階層に属する遺伝子、酵素ＥＣＡと同一グループに属する遺伝子もしくは前記いずれかの遺伝子の組を含む条件とする。

【００５４】

表４から表６において遺伝子間の相互関係が階層構造でかつスーパーファミリーであった場合、クラス分類関連抽出遺伝子のスーパーファミリーがＳＦＡであれば、アノテーションから抽出された共通性や規則性に基づく拘束条件を同一スーパーファミリーＳＦＡに属する遺伝子、もしくは前記の抽出された遺伝子と前記の同一スーパーファミリーの属する遺伝子との組を含む条件とする。

【００５５】

表４から表６において遺伝子間の相互関係がネットワークでかつ文献情報であった場合、アノテーションから抽出された共通性や規則性に基づく拘束条件とは、文献情報により、クラス分類関連抽出遺伝子との関連が予想される遺伝子、もしくはクラス分類関連抽出遺伝子と前記の文献情報から関連が予想される遺伝子との組を含む条件とする。

【００５６】

表４から表６において遺伝子間の相互関係がネットワークでかつ蛋白質相互作用であった場合、アノテーションから抽出された共通性や規則性に基づく拘束条件とは、蛋白質相互作用により、クラス分類関連抽出遺伝子との関連が予想される遺伝子、もしくはクラス分類関連抽出遺伝子と前記の蛋白質相互作用から関連が予想される遺伝子との組を含む条件とする。

【００５７】

表４から表６において遺伝子間の相互関係がネットワークでかつ代謝経路情報であった場合、アノテーションから抽出された共通性や規則性に基づく拘束条件とは、代謝経路情報により、クラス分類関連抽出遺伝子との関連が予想される遺伝子、もしくはクラス分類関連抽出遺伝子と前記の代謝経路情報から関連が予想される遺伝子との組を含む条件とする。

本発明において、図１４、図１５、表４から６に開示した方法を用いて反復マイニングを行うことで、より一般性や頑強性（ロバストネス）が高いメタ知識（ｍｅｔａ－ｋｎｏｗｌｅｄｇｅ）を自動で得ることができる。

【００５８】

【発明の実施の形態】

公開データである急性白血病のデータセット（Ｇｏｌｕｂら１９９９年）を使用し、具体的な発明の実施形態を示す。データファイルはＭＩＴのサイトからﾀﾞｳﾝﾛｰﾄﾞした（ｈｔｔｐ：／／ｗｗｗ．ｇｅｎｏｍｅ．ｗｉ．ｍｉｔ．ｅｄｕ／ＭＰＲ）。このデータセットは、２種類の急性白血病：ＡＬＬ（ａｃｕｔｅ　ｌｙｍｐｈｏｂｌａｓｔｉｃ　ｌｅｕｋｅｍｉａ）、ＡＭＬ（ａｃｕｔｅ　ｍｙｅｌｏｉｄ　ｌｅｕｋｅｍｉａ）患者から血液を採取し、Ａｆｆｙｍｅｔｒｉｘヒトチップ（６８１７遺伝子）を用いて発現分布を測定したものである。被験者数は、合計７２人（内、３８人がＢ－ｃｅｌｌ　ＡＬＬ、９人がＴ－ｃｅｌｌ　ＡＬＬ、２５人がＡＭＬ）である。このデータセット（７２×６８１７）を用いて、発現データのみから２種類の白血病をクラシフィケーション（分類）することができるかを示す。以下、２種類の白血病（ＡＬＬ，　ＡＭＬ）をそれぞれ、クラス０，１と定義する。すなわち、ＡＬＬ＝クラス０、ＡＭＬ＝クラス１とする。なお急性白血病のデータセット（Ｇｏｌｕｂら１９９９年）は、７２検体データを３８のトレーニングセット（ｄａｔａ＿ｓｅｔ＿ＡＬＬ＿ＡＭＬ＿ｔｒａｉｎ．ｔｘｔ）と、３４のテストセット（ｄａｔａ＿ｓｅｔ＿ＡＬＬ＿ＡＭＬ＿ｉｎｄｅｐｅｎｄｅｎｔ．ｔｘｔ）に分けて提供されている。そのため、ここでは３８検体（トレーニングセット）を用いて学習をおこない、その結果を用いて３４検体（テストセット）のクラシフィケーションを試みた。この分類の結果と、予め分かっている各検体のクラスとを比較することで正解率とエラー率を求めることができる。

【００５９】

以後特徴ルール法を用いてデータマイニングした例を示す。６８１８属性、７２サンプル

のデータセットを対象データとした。このうちの６８１７属性は、ＤＮＡチップで測定した遺伝子１つ１つの発現量に対応し、残りの１属性は、白血病の種類（ＡＬＬ、ＡＭＬ）に対応している。遺伝子発現量に対応する属性値は実数値データであるため、それぞれの属性項目ごとに「大」「中」「小」の３カテゴリに離散化した。この時、７２個のサンプルが各カテゴリにほぼ均等に振り分けられるようにそれぞれの属性項目のカテゴリの境界を設定した。また、白血病の種類に対応する属性値はもともと２種類の離散値（クラス０、クラス１）なので、そのまま用いた。

【００６０】

遺伝子に対応する６８１７属性をＩＦ－ＴＨＥＮルールの条件項目、白血病の種類に対応する属性を結論項目、「クラス０」を結論項目値とし、また、条件部に現れる述語数の上限は１として、評価値の上位２０個のルールを取り出した結果を表７に示す。表７の各行が１つのＩＦ－ＴＨＥＮルールに対応しており、評価値の大きい順に上から並んでいる。表７の第１列は、ＩＦ－ＴＨＥＮルールの条件部に対応し、第２列、第３列、第４列はそれぞれ、ルールの評価値、ヒット率、カバー率に対応する。なお、ルールの結論部は全てのルールで「白血病の種類　＝　クラス０」で同一なので、表中では省略した。例えば、第１行は「ＩＦ　Ｕ０７１３９＿ａｔ　＝　大　ＴＨＥＮ　白血病の種類　＝　クラス０」というルールであり、「Ｕ０７１３９＿ａｔ」という遺伝子の発現量が「大」であるならば、白血病の種類はクラス０（ＡＬＬ）となる傾向が大きいということを意味している。

【００６１】

ところで、第１、第２のルールの評価値はともに　０．１４８　で同じ、第３のルール以降は全て評価値　０．１４２　で同じである。このような評価値の同じルール同士の並び順には分析上の意味はなく、データファイルの中の属性項目の並び順に依存している。

【００６２】

【表７】

表７　特徴ルール法を用いたデータマイニング例

条件部	ルール評価値	セット率	カバー率
U07139_at＝大	0.148	1.0	0.35
M94633_at＝大	0.148	1.0	0.35
U29175_at＝大	0.142	1.0	0.33
M28170_at＝大	0.142	1.0	0.33
U27460_at＝大	0.142	1.0	0.33
M31523_at＝大	0.142	1.0	0.33
X97267_rna1_s_at＝大	0.142	1.0	0.33
X85116_rna1_s_at＝小	0.142	1.0	0.33
M84371_rna1_s_at＝大	0.142	1.0	0.33
M12959_s_at＝大	0.142	1.0	0.33
L09209_s_at＝小	0.142	1.0	0.33
U22376_cds2_s_at＝大	0.142	1.0	0.33
D26156_s_at＝大	0.142	1.0	0.33
Z49194_at＝大	0.142	1.0	0.33
Y08612_at＝大	0.142	1.0	0.33
X99920_at＝大	0.142	1.0	0.33
X95735_at＝小	0.142	1.0	0.33
X93512_at＝大	0.142	1.0	0.33
X82240_rna1_at＝大	0.142	1.0	0.33
X68560_at＝大	0.142	1.0	0.33

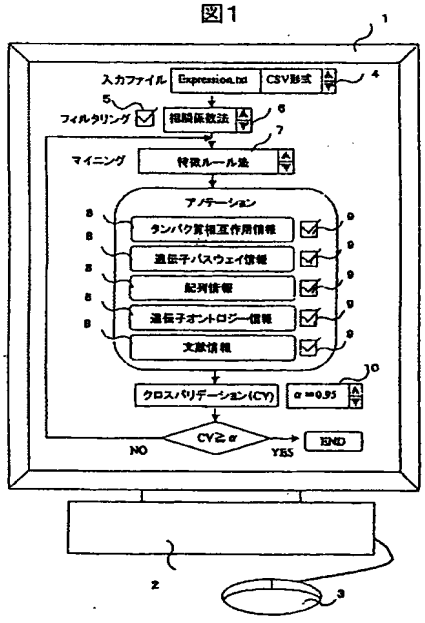
【発明の効果】
本発明により、遺伝子発現データに基づく、知識探索を行うことができる。遺伝子等同士の相互作用アノテーションに基づく繰り返し探索を行うことで、より一般性や頑強性（ロバストネス）が高いメタ知識（ｍｅｔａ－ｋｎｏｗｌｅｄｇｅ）を自動で得ることができる。

【図面の簡単な説明】
【図１】本発明の実施形態のソフトウェアを実行するために利用されるコンピュータシステムとその操作画面例。
【図２】ＤＮＡチップの概要図。
【図３】ゲノム、トランスクリプトーム、プロテオームにおける相互作用の例。
【図４】二項関係の一例であるリガンドーレセプタ関係の例。
【図５】パスウェイ関係の例。
【図６】ゲノム関係の例。
【図７】階層構造の一例である遺伝子オントロジー関係の例。
【図８】階層構造の一例である酵素（ＥＣ：Ｅｎｚｙｍｅ　Ｃｏｍｍｉｓｓｉｏｎ）関係の例。
【図９】階層構造の一例であるスーパーファミリー関係の例。
【図１０】ネットワークの一例である文献情報の例。
【図１１】ネットワークの一例である遺伝子相互作用の例。
【図１２】ネットワークの一例である代謝経路の例。
【図１３】生物情報の階層構造の例。

【図１４】クロスバリデーション値と既定値とを比較することで繰り返し回数を決定する、本発明を実現するためのコンピュータ化された方法を示す模式図である。
【図１５】全てのクロスバリデーション値を保存して比較する、本発明を実現するためのコンピュータ化された方法を示す模式図である。
【図１６】ＤＮＡチップを用いた測定法のフローチャートの一例である。
【図１７】発現マトリクスの模式図である。
【符号の説明】
１．コンピュータシステム、２．ＣＰＵ、３．入力装置（マウス）、４．入力ファイル名・形式選択、５．フィルタリング機能オンオフボタン、６．フィルタリングアルゴリズム選択、７．マイニング手法選択、８．アノテーション内容、９．アノテーション内容選択、１０．イタレーション条件選択、２１．蛍光検出器、２２．ＤＮＡプローブ、２３．蛍光標識された遺伝子、２４．支持体、４１．リガンド、４２．レセプター、５１．遺伝子、５２．遺伝子間パスウェイ、６１．染色体マップ、６２、遺伝子マップ、６３．遺伝子名、７１．遺伝子オントロジー（ＤＮＡ ｒｅｐａｉｒ）、７２．遺伝子オントロジー（ｐｒｏｔｅｉｎａｍｉｎｏ ａｃｉｄ ＡＤＰ－ｒｉｂｏｓｙｌａｔｉｏｎ）、８１．酵素（ＥＣ）、９１．遺伝子スーパーファミリー、１０１．遺伝子、１０２．相互関係スコア、１１１．蛋白質、１１２．蛋白質間相互作用、１２１．酵素、１２２．代謝パスウェイ、１２３．酵素反応生成物、１７１．サンプルアノテーション、１７２．遺伝子アノテーション、１７３．発現レベル。

【図１】

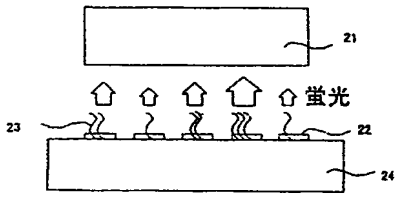
図1



【図２】

図2

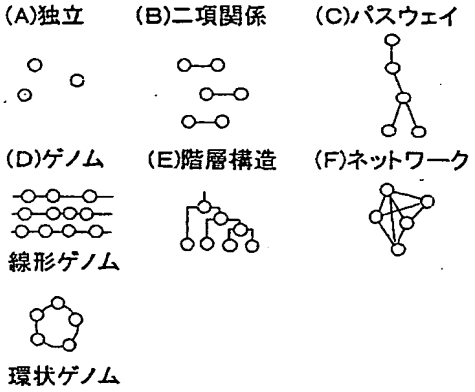
DNAチップの概念図



【図３】

図3

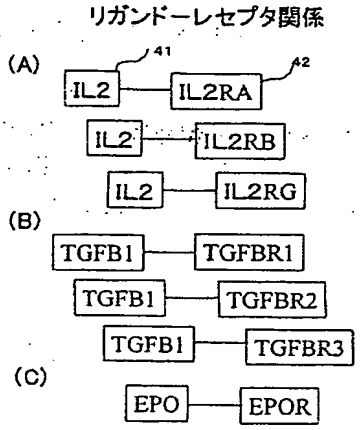
ゲノム、トランスクリプトーム、プロテオーム
における相互関係の例



【図４】

図4

二項関係の例

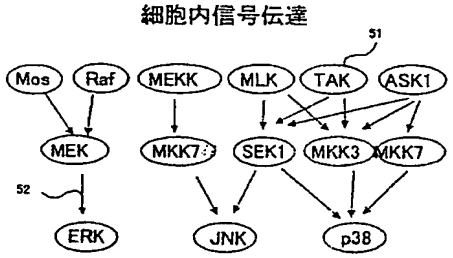


【図５】

図5

パスウェイの例

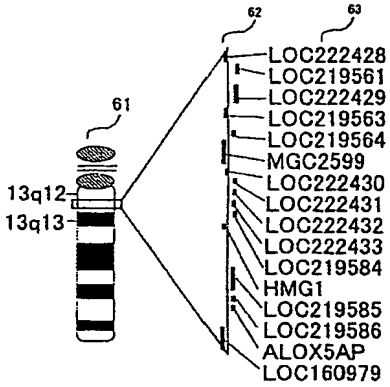
細胞内信号伝達



【図６】

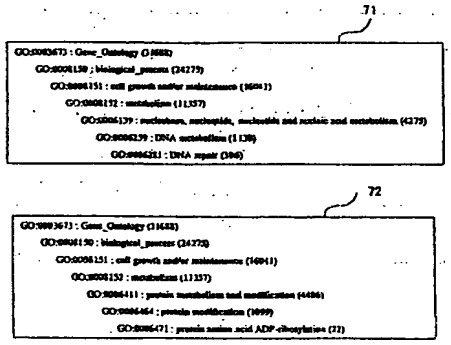
図6

ゲノムの例



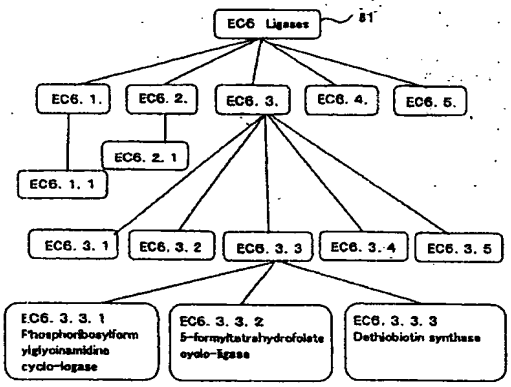
【図７】

図7
階層構造の例(その1)
遺伝子オントロジー



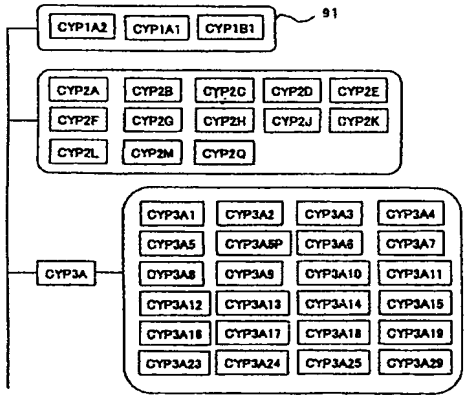
【図８】

図8
階層構造の例(その2)
酵素(EC:Enzyme Commission)



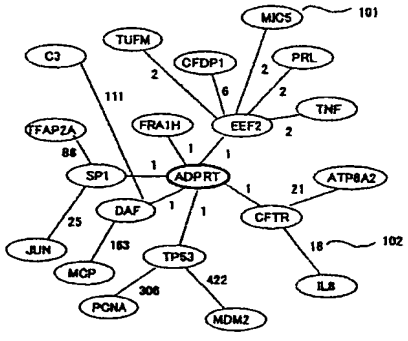
【図９】

図9
階層構造の例(その3)
スーパーファミリー



【図１０】

図10
ネットワークの例(その1)
文献情報

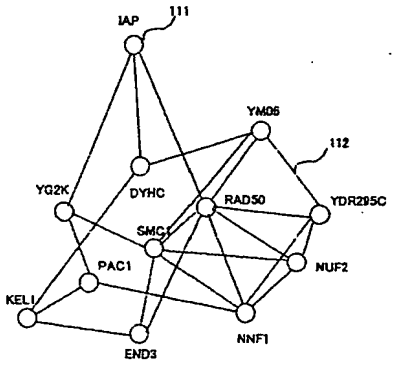


【図１１】

図11

ネットワークの例(その2)

蛋白質相互作用

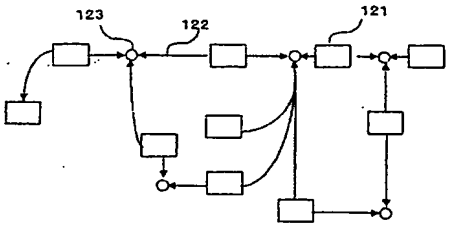


【図１２】

図12

ネットワークの例(その3)

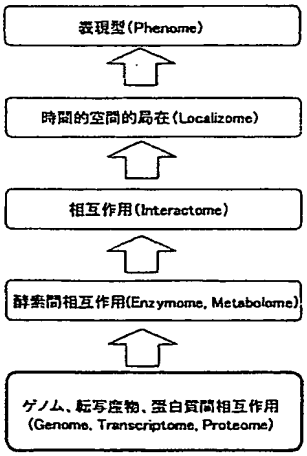
代謝経路



【図１３】

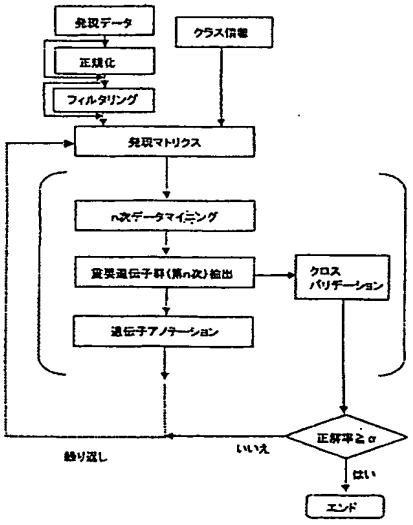
図13

生物情報の階層構造の例



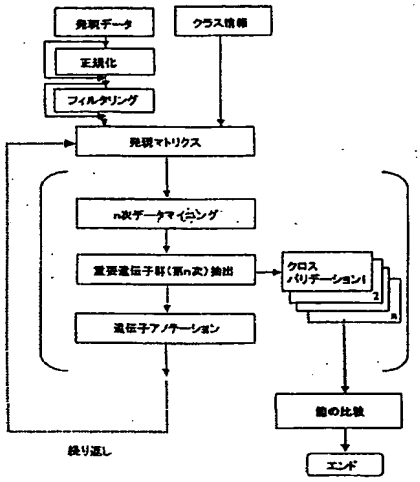
【図１４】

図14



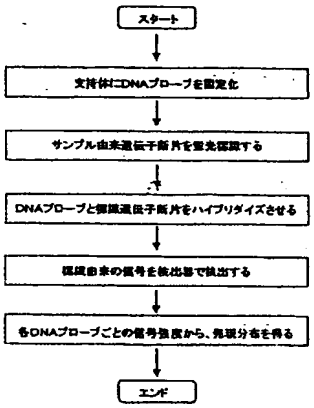
【図１５】

図15



【図１６】

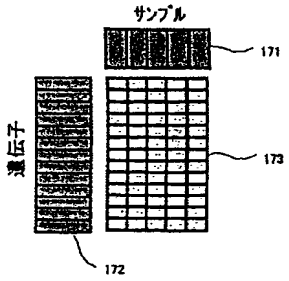
図16



【図１７】

図17

発現マトリクス



フロントページの続き

(72)発明者 森田 豊久
神奈川県川崎市麻生区王禅寺１０９９番地 株式会社日立製作所システム開発研究所内
(72)発明者 ソリン サバウ
東京都千代田区神田駿河台四丁目６番地 株式会社日立製作所ライフサイエンス推進事業部内
(72)発明者 谷川 浩司
東京都千代田区神田駿河台四丁目６番地 株式会社日立製作所ライフサイエンス推進事業部内

Fターム(参考) 5B075 ND20 NS10 UU19